



Predicting Harmful
Algal Blooms

The Problem

Toxins produced by Harmful Algal Blooms (HABs) are a health hazard to humans and animals, and incur a large financial burden to water utilities and other water quality stakeholders



Algae like
cyanobacteria
can create toxins



The Problem

Toxins produced by Harmful Algal Blooms (HABs) are a health hazard to humans and animals, and incur a large financial burden to water utilities and other water quality stakeholders



Algae like cyanobacteria can create toxins



Just like the weather, value is in knowing **IN ADVANCE** that a bloom is coming

The Problem

Toxins produced by Harmful Algal Blooms (HABs) are a health hazard to humans and animals, and incur a large financial burden to water utilities and other water quality stakeholders



Algae like cyanobacteria can create toxins



Just like the weather, value is in knowing **IN ADVANCE** that a bloom is coming

So we collect data



Lots of data: water samples, CYAN, qPCR...etc

The Problem

Toxins produced by Harmful Algal Blooms (HABs) are a health hazard to humans and animals, and incur a large financial burden to water utilities and other water quality stakeholders



Algae like cyanobacteria can create toxins

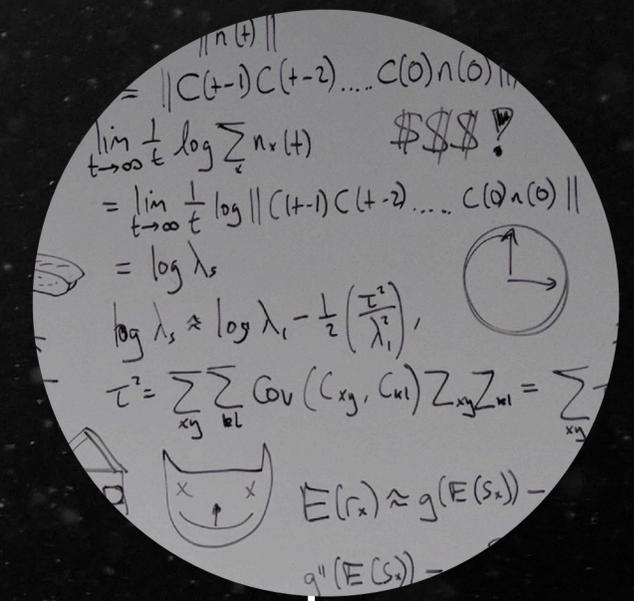


Just like the weather, value is in knowing **IN ADVANCE** that a bloom is coming

So we collect data



Lots of data: water samples, CYAN, qPCR...etc



Then we use math and algorithms, but...

What's The Real Problem?

We collect a lot of data, we develop a lot of models... but how do we make the data we have actionable?



What's The Real Problem?

We collect a lot of data, we develop a lot of models... but how do we make the data we have actionable?

Given identification of the drivers of HABs, can we reduce their frequency?



What's The Real Problem?

We collect a lot of data, we develop a lot of models... but how do we make the data we have actionable?

Given identification of the drivers of HABs, can we reduce their frequency?

If a HAB is going to happen, how best can we prepare?



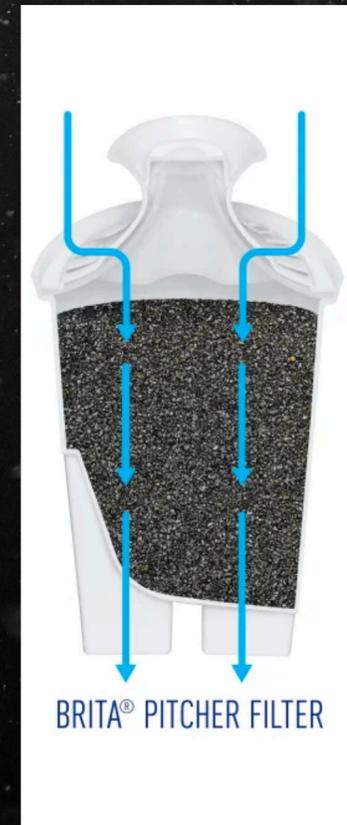
What's The Real Problem?

We collect a lot of data, we develop a lot of models... but how do we make the data we have actionable?

Given identification of the drivers of HABs, can we reduce their frequency?

If a HAB is going to happen, how best can we prepare?

Can we stop HABs once they've started?



What's The Real Problem?

We collect a lot of data, we develop a lot of models... but how do we make the data we have actionable?

Given identification of the drivers of HABs, can we reduce their frequency?

If a HAB is going to happen, how best can we prepare?

Can we stop HABs once they've started?



Challenges

Our algorithms focus on forecasting HAB occurrence: These are the major challenges we've faced:



Challenges

Our algorithms focus on forecasting HAB occurrence: These are the major challenges we've faced:



Ecosystem
Uniqueness
(no general
solution
possible)



Challenges

Our algorithms focus on forecasting HAB occurrence: These are the major challenges we've faced:



Ecosystem
Uniqueness
(no general
solution
possible)



Differential
Sampling
(data rich vs
data poor)



Challenges

Our algorithms focus on forecasting HAB occurrence: These are the major challenges we've faced:



Ecosystem
Uniqueness
(no general
solution
possible)



Differential
Sampling
(data rich vs
data poor)



Ecosystem
Connectivity
(open vs
closed
systems)

Challenges

Our algorithms focus on forecasting HAB occurrence: These are the major challenges we've faced:



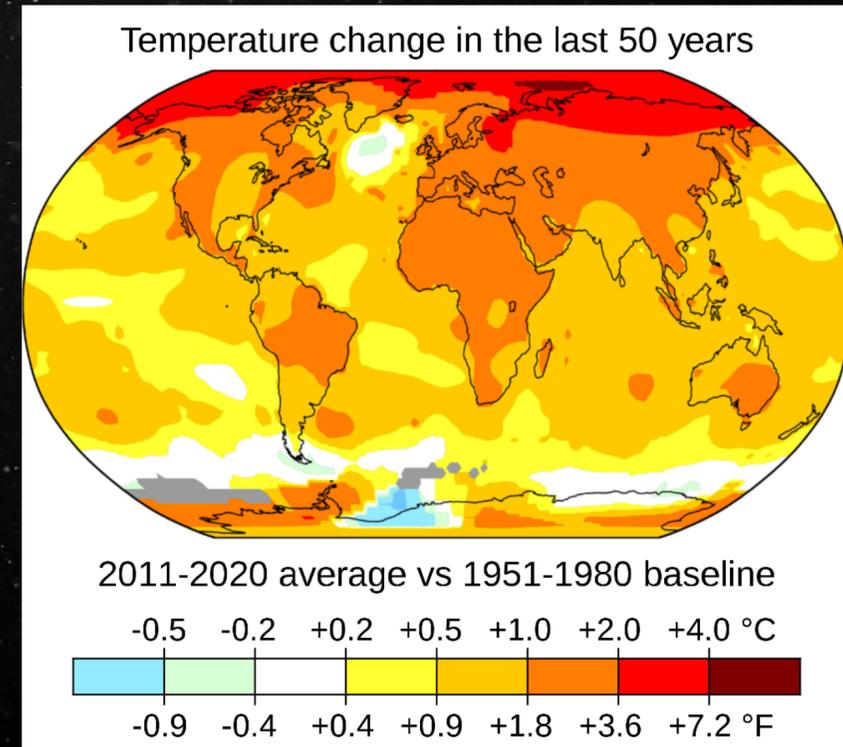
Ecosystem
Uniqueness
(no general
solution
possible)



Differential
Sampling
(data rich vs
data poor)



Ecosystem
Connectivity
(open vs
closed
systems)



Timescales
(weeks, months,
years, decades)



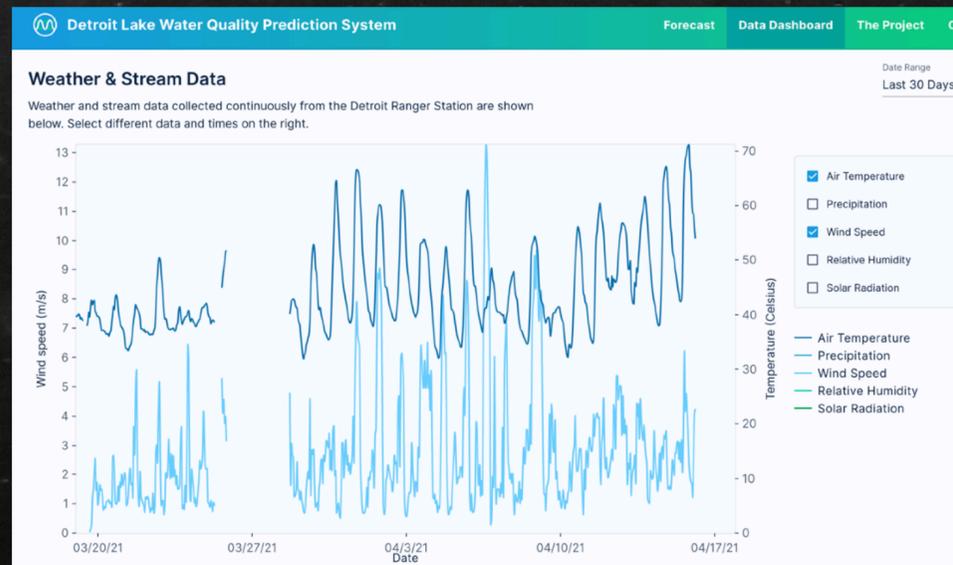
Solutions

We have been developing cyberinfrastructure and a suite of data modeling tools that help address these HAB challenges:



Solutions

We have been developing cyberinfrastructure and a suite of data modeling tools that help address these HAB challenges:



Operational
(Detroit Lake)

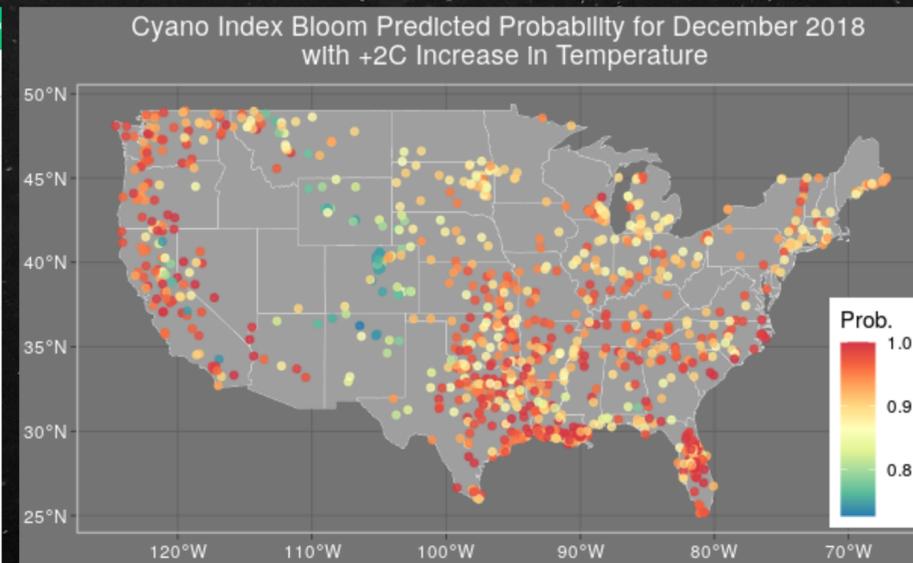


Solutions

We have been developing cyberinfrastructure and a suite of data modeling tools that help address these HAB challenges:



Operational
(Detroit Lake)



Seasonal and
Climate sensitivity
(CYAN)

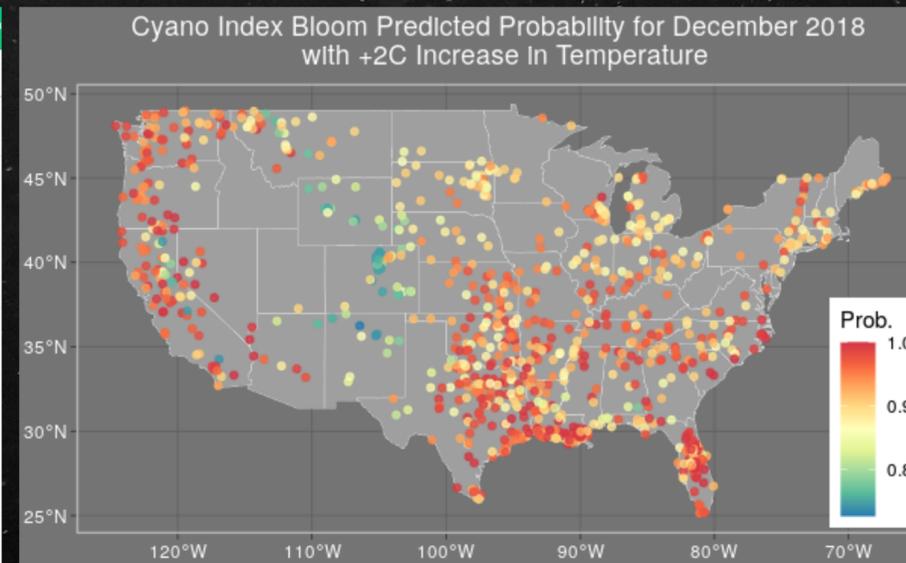


Solutions

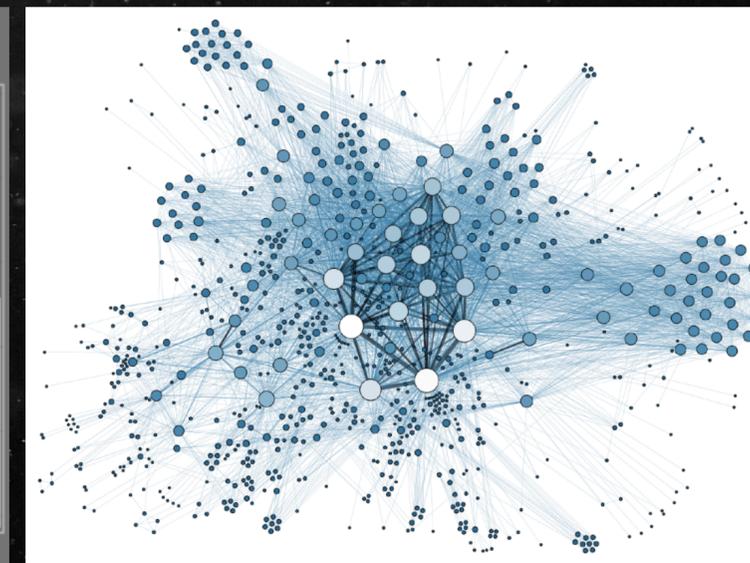
We have been developing cyberinfrastructure and a suite of data modeling tools that help address these HAB challenges:



Operational
(Detroit Lake)



Seasonal and
Climate sensitivity
(CYAN)



Next Gen
Analytics

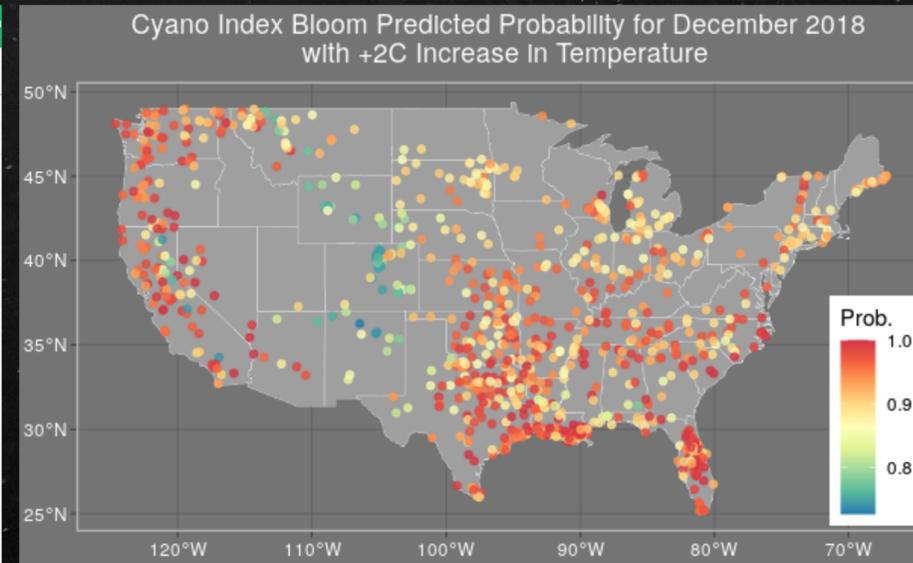


Solutions

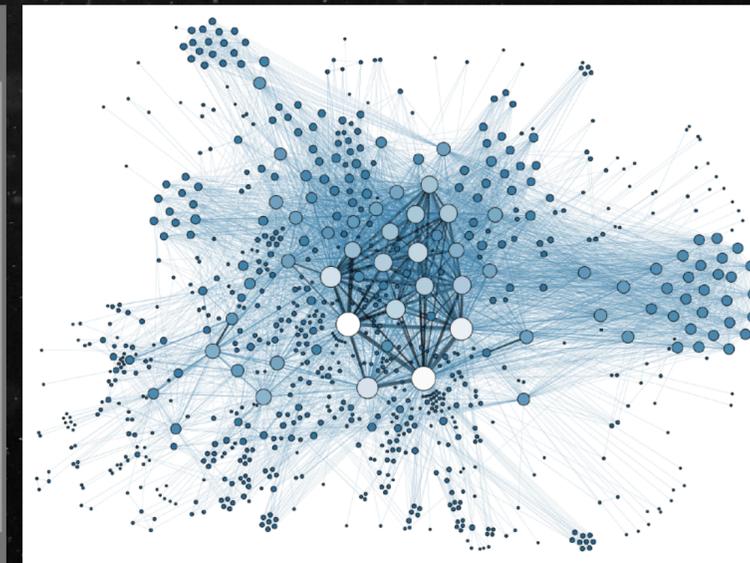
We have been developing cyberinfrastructure and a suite of data modeling tools that help address these HAB challenges:



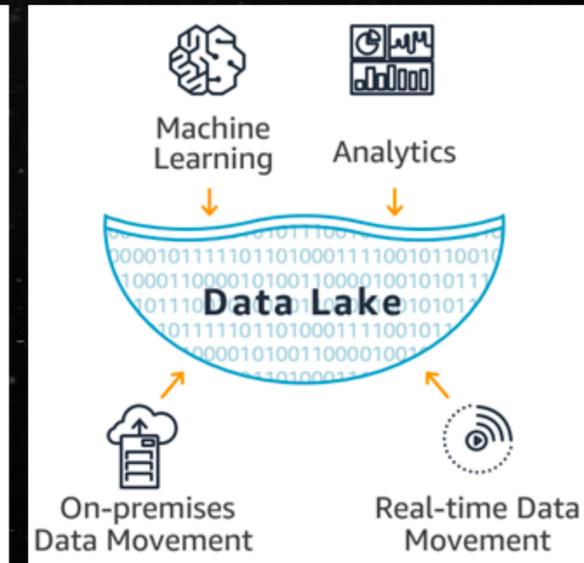
Operational
(Detroit Lake)



Seasonal and
Climate sensitivity
(CYAN)



Next Gen
Analytics



Cloud
Infrastructure



Operational Forecasts

Detroit Lake: Bayesian Model Averaging framework applied to the in situ data collected from the lake to provide daily 1-week and 2-week forecasts of cyanobacteria and toxin concentrations



Operational Forecasts

Detroit Lake: Bayesian Model Averaging framework applied to the in situ data collected from the lake to provide daily 1-week and 2-week forecasts of cyanobacteria and toxin concentrations

Ecological Applications, 19(7), 2009, pp. 1805–1814
© 2009 by the Ecological Society of America

Bayesian model averaging for harmful algal bloom prediction

GRANT HAMILTON,^{1,4} ROSS McVINISH,² AND KERRIE MENGENSEN³

¹*School of Natural Resource Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, Queensland 4001 Australia*

²*Mathematics Department, The University of Queensland, Brisbane, Queensland 4072 Australia*

³*School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, Queensland 4001 Australia*

We expanded the Bayesian Model Averaging approach to include neural nets in addition to linear and non-linear functions



Based on empirical samples



Operational Forecasts

Detroit Lake: Bayesian Model Averaging framework applied to the in situ data collected from the lake to provide daily 1-week and 2-week forecasts of cyanobacteria and toxin concentrations

Ecological Applications, 19(7), 2009, pp. 1805–1814
© 2009 by the Ecological Society of America

Bayesian model averaging for harmful algal bloom prediction

GRANT HAMILTON,^{1,4} ROSS McVINISH,² AND KERRIE MENGERSEN³

¹*School of Natural Resource Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, Queensland 4001 Australia*

²*Mathematics Department, The University of Queensland, Brisbane, Queensland 4072 Australia*

³*School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, Queensland 4001 Australia*

We expanded the Bayesian Model Averaging approach to include neural nets in addition to linear and non-linear functions



Based on empirical samples



Operational Dashboard
(Detroit Lake)

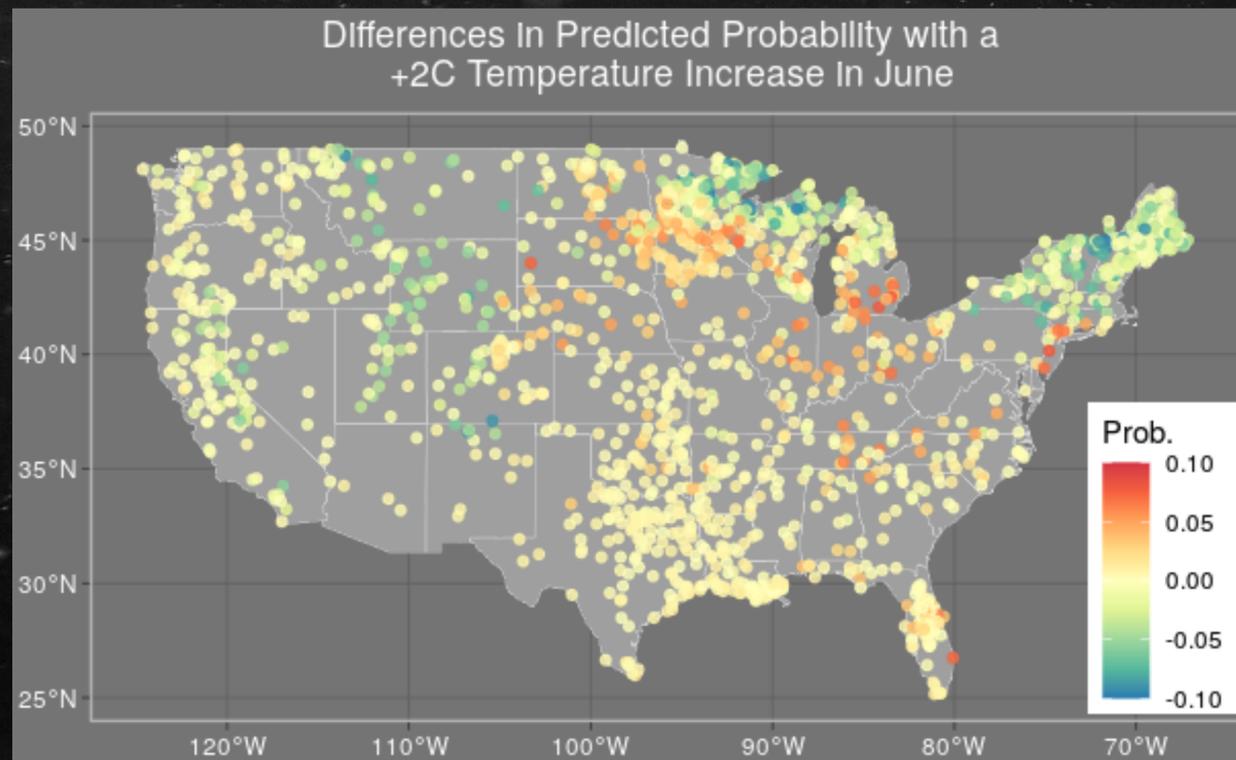


Long-term Forecasts

In addition to operational forecasts (1-2 week timescales), we have been developing **seasonal** (6 month) and **decadal** forecasts (e.g. 2030, 2050, 2100).

Long-term Forecasts

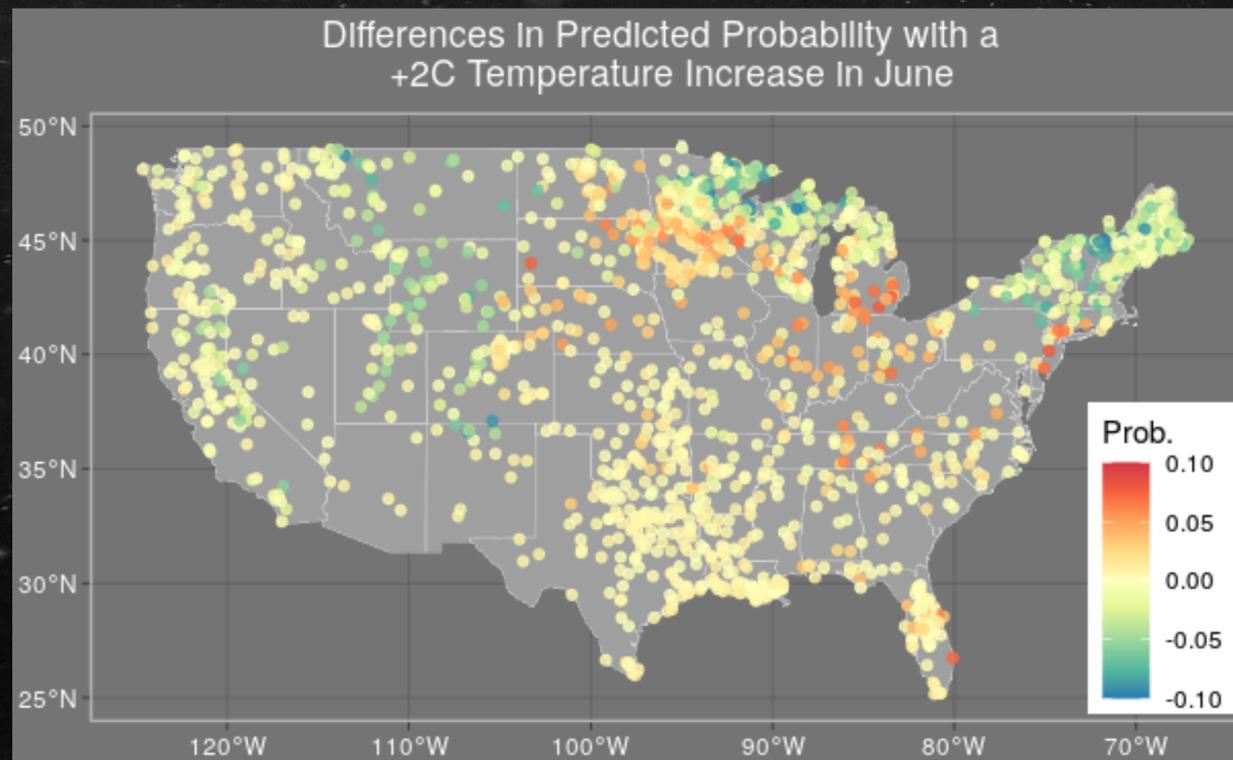
In addition to operational forecasts (1-2 week timescales), we have been developing **seasonal** (6 month) and **decadal** forecasts (e.g. 2030, 2050, 2100).



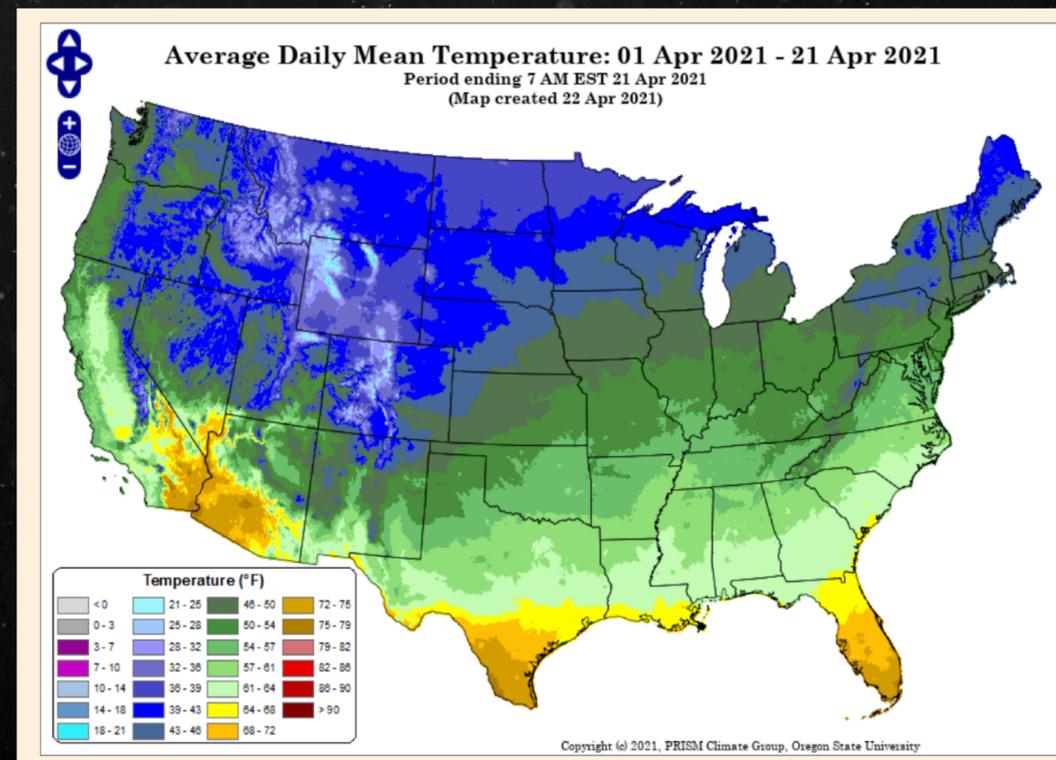
CYAN national data

Long-term Forecasts

In addition to operational forecasts (1-2 week timescales), we have been developing **seasonal** (6 month) and **decadal** forecasts (e.g. 2030, 2050, 2100).



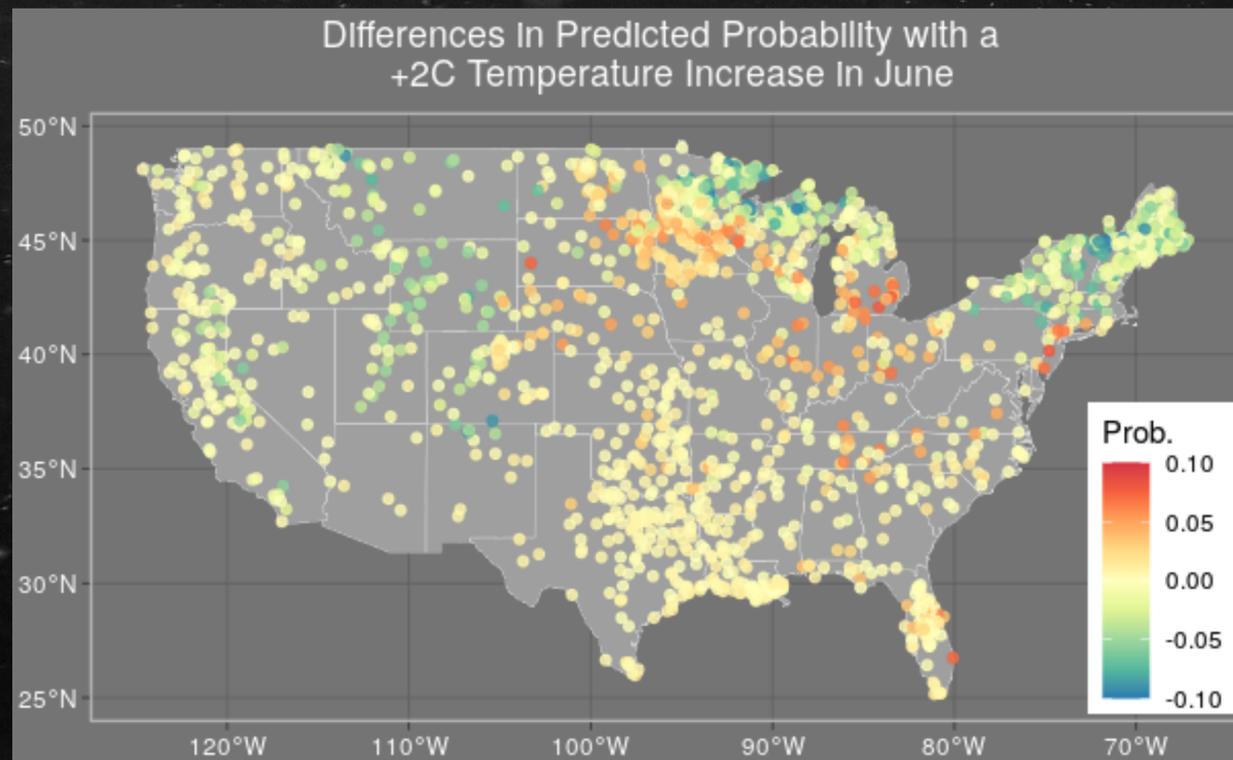
CYAN national data



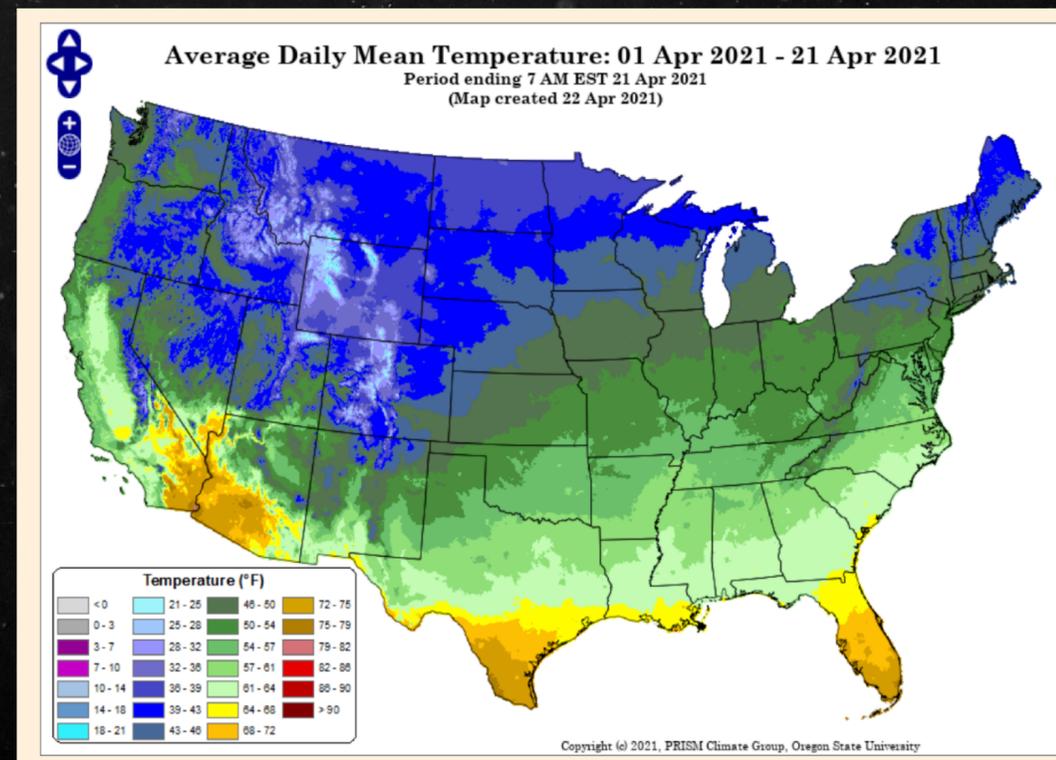
PRISM (OSU) downscaled weather/climate data

Long-term Forecasts

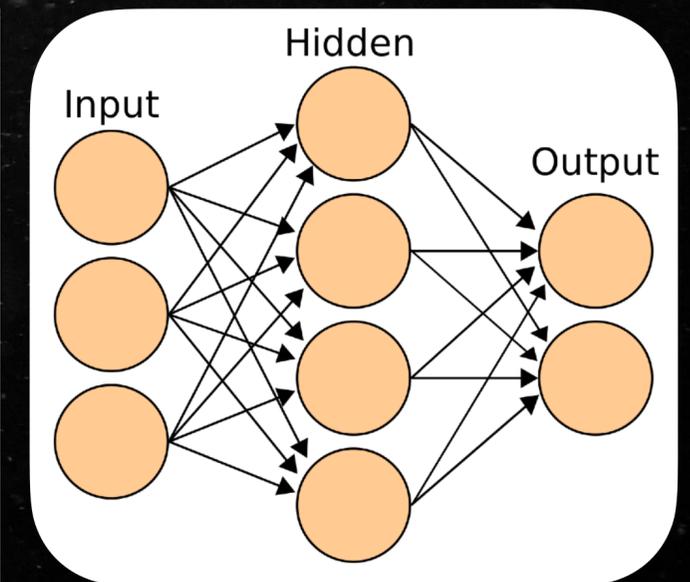
In addition to operational forecasts (1-2 week timescales), we have been developing **seasonal** (6 month) and **decadal** forecasts (e.g. 2030, 2050, 2100).



CYAN national data



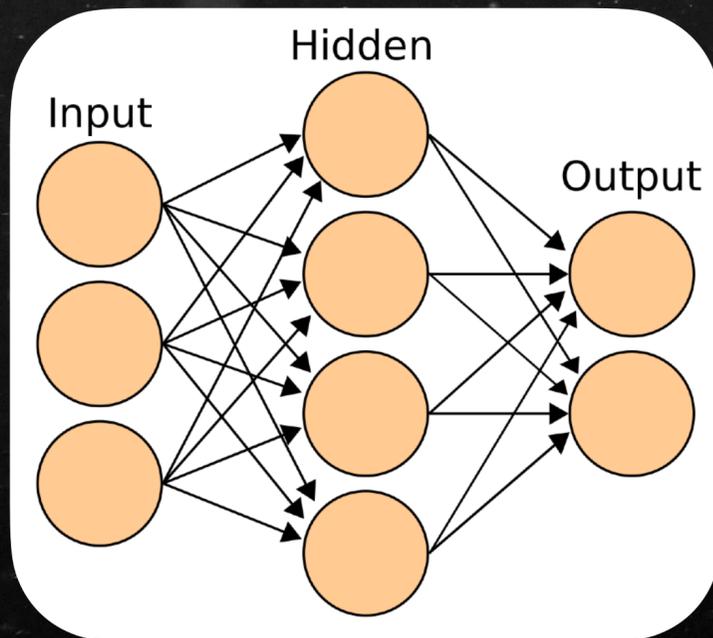
PRISM (OSU) downscaled weather/climate data



Machine learning modeling

Long-term Forecasts

In addition to operational forecasts (1-2 week timescales), we have been developing **seasonal** (6 month) and **decadal** forecasts (e.g. 2030, 2050, 2100).

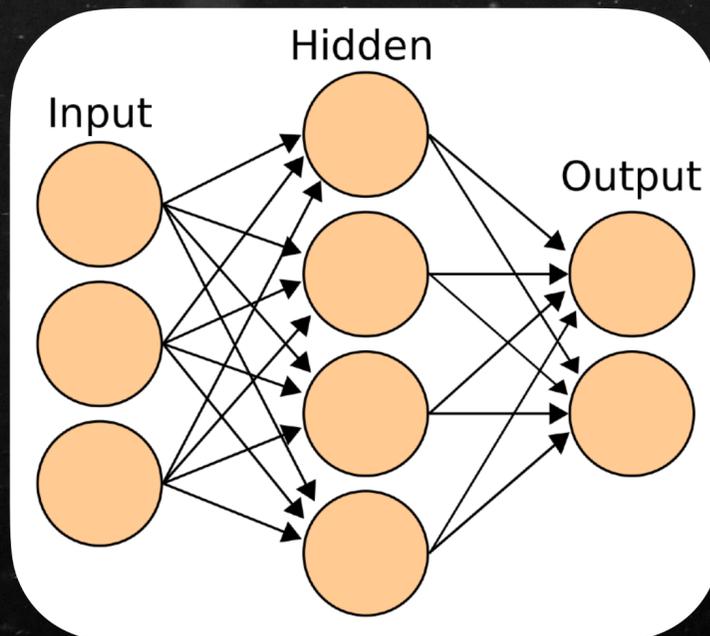


Machine learning
modeling

Long-term Forecasts

In addition to operational forecasts (1-2 week timescales), we have been developing **seasonal** (6 month) and **decadal** forecasts (e.g. 2030, 2050, 2100).

Long-range weather forecasts



Machine learning modeling

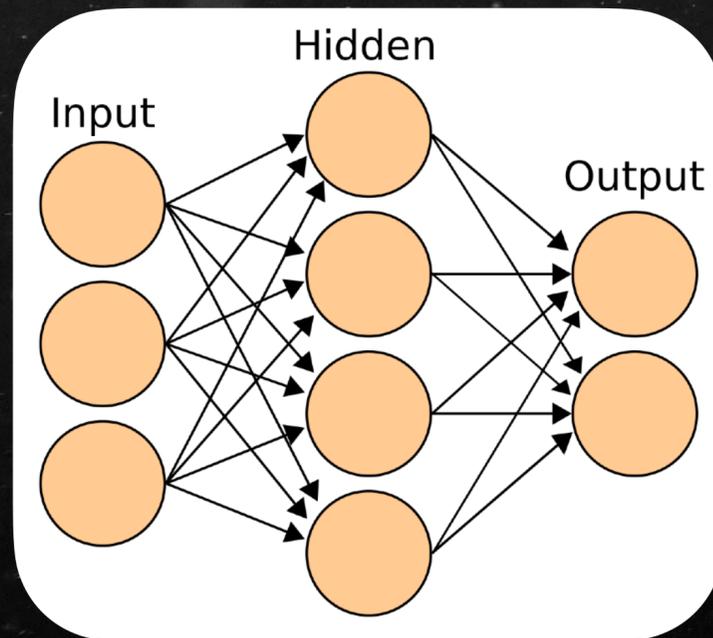
A screenshot of the National Weather Service website. The header includes the NOAA logo and the text "NATIONAL WEATHER SERVICE NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION". A navigation menu contains links for HOME, FORECAST, PAST WEATHER, SAFETY, INFORMATION, EDUCATION, NEWS, SEARCH, and ABOUT. Below the menu, there is a search box for "Local forecast by 'City, St' or ZIP code" with a "Go" button and a "Location Help" link. The main content area features a headline: "Severe Storms Possible in the South; Southwest Critical Fire Weather Threats". The text below the headline reads: "Severe storms with large to very large hail, damaging winds, a couple of tornadoes, and heavy rain that could cause flash flooding may develop across the eastern South Plains to the Deep South. Dry, gusty winds in the Southwest will keep fire weather threats critical. Elevated fire weather threats also in portions of the East. Cool and snowy in the Northern Tier states and West mountains." A "Read More >" link is provided at the bottom of the text.



Long-term Forecasts

In addition to operational forecasts (1-2 week timescales), we have been developing **seasonal** (6 month) and **decadal** forecasts (e.g. 2030, 2050, 2100).

Long-range weather forecasts



Machine learning modeling

A screenshot of the National Weather Service website. The header includes the NOAA logo and the text "NATIONAL WEATHER SERVICE NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION". A navigation menu contains links for HOME, FORECAST, PAST WEATHER, SAFETY, INFORMATION, EDUCATION, NEWS, SEARCH, and ABOUT. Below the menu, there is a search bar for "Local forecast by 'City, St' or ZIP code" with a "Go" button and a "Location Help" link. The main content area features a headline: "Severe Storms Possible in the South; Southwest Critical Fire Weather Threats". Below the headline, a paragraph of text describes the forecast: "Severe storms with large to very large hail, damaging winds, a couple of tornadoes, and heavy rain that could cause flash flooding may develop across the eastern South Plains to the Deep South. Dry, gusty winds in the Southwest will keep fire weather threats critical. Elevated fire weather threats also in portions of the East. Cool and snowy in the Northern Tier states and West mountains." A "Read More >" link is provided at the bottom of the text.

A screenshot of the WCRP (World Climate Research Programme) website. The header features the WCRP logo and the text "World Climate Research Programme". Below the logo, there are logos for the World Meteorological Organization, the United Nations Educational, Scientific and Cultural Organization, and the Intergovernmental Oceanographic Commission. The International Science Council logo is also present. A navigation menu at the bottom includes links for Home, About WCRP, Core Projects, Unifying Themes, Grand Challenges, Initiatives & Activities, and Events.

CMIP6 Climate Projections



Next Gen: Hybrid Machine Learning

Purely data driven machine learning (e.g. our Bayesian Model Averaging, or Neural Nets, or Random Forests) **do not explicitly respect the laws of physics or biology.**

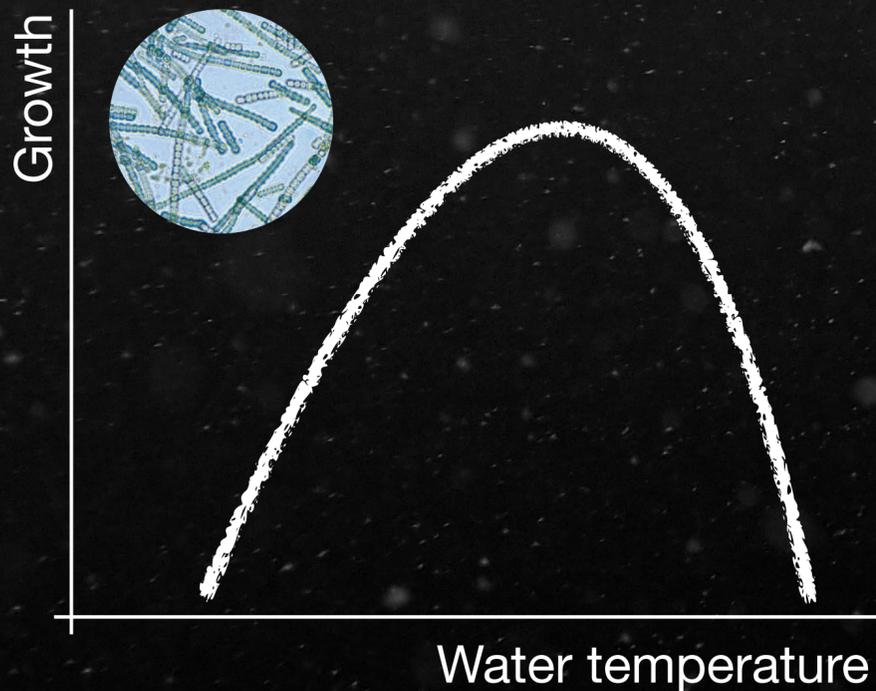
But, hybrid machine learning models do. Case study - this year!



Next Gen: Hybrid Machine Learning

Purely data driven machine learning (e.g. our Bayesian Model Averaging, or Neural Nets, or Random Forests) **do not explicitly respect the laws of physics or biology.**

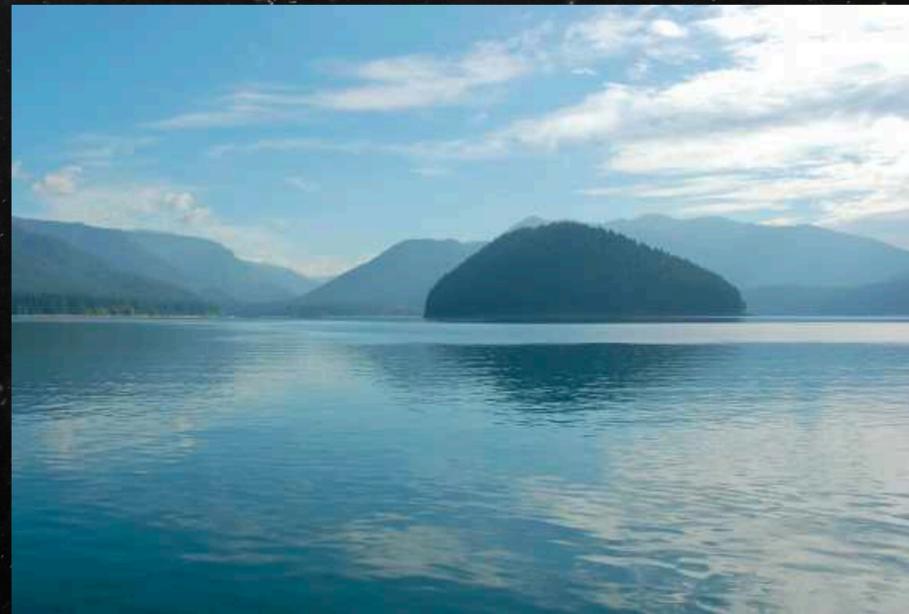
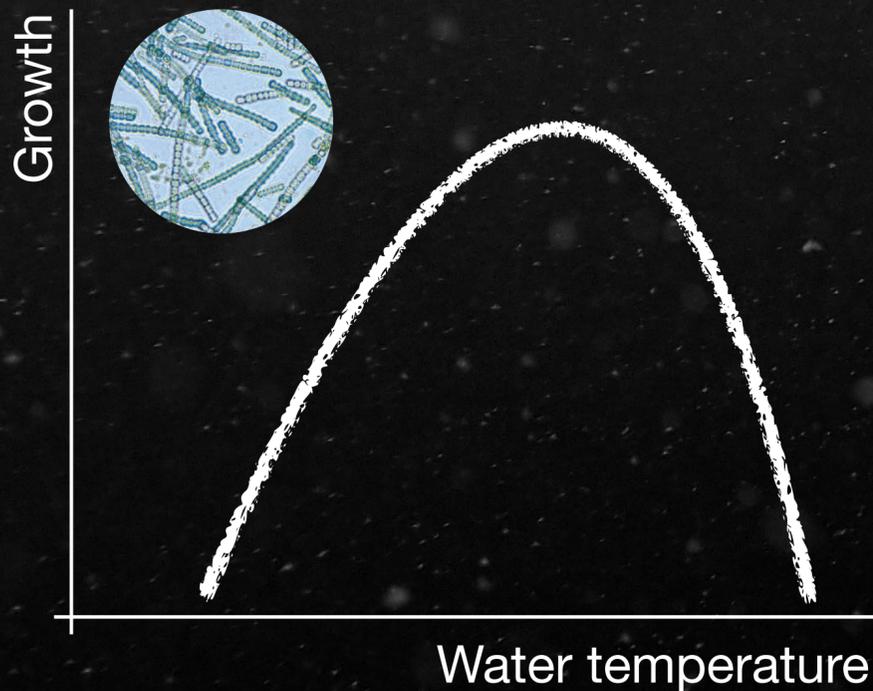
But, hybrid machine learning models do. Case study - this year!



Next Gen: Hybrid Machine Learning

Purely data driven machine learning (e.g. our Bayesian Model Averaging, or Neural Nets, or Random Forests) **do not explicitly respect the laws of physics or biology.**

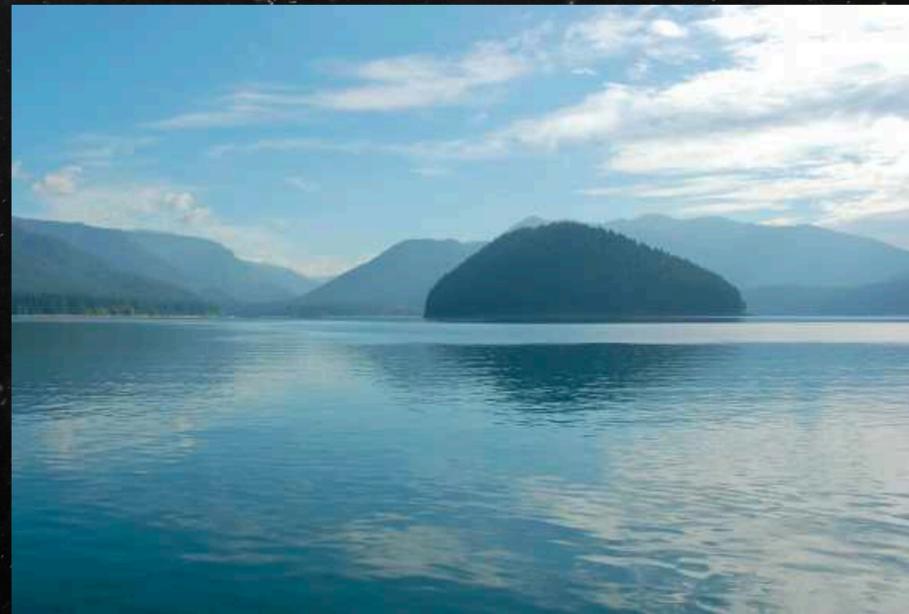
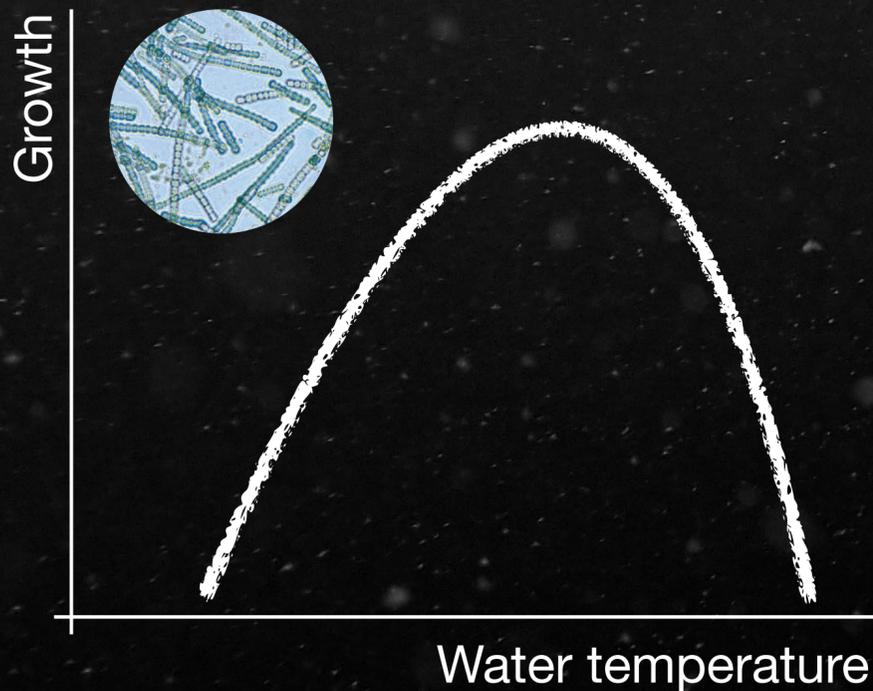
But, hybrid machine learning models do. Case study - this year!



Next Gen: Hybrid Machine Learning

Purely data driven machine learning (e.g. our Bayesian Model Averaging, or Neural Nets, or Random Forests) **do not explicitly respect the laws of physics or biology.**

But, hybrid machine learning models do. Case study - this year!



Next Gen: Interpretable AI

Machine learning (e.g. neural nets) offer accurate predictions, but you don't know what's going on under the hood. **Interpretable AI** is a class of machine learning that people can understand

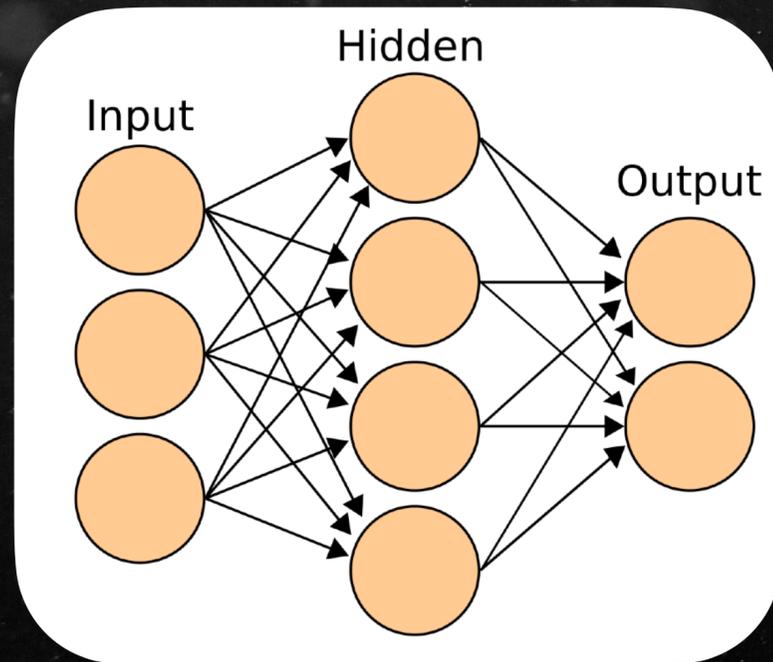
The Learning Boundary



Next Gen: Interpretable AI

Machine learning (e.g. neural nets) offer accurate predictions, but you don't know what's going on under the hood. **Interpretable AI** is a class of machine learning that people can understand

Black box
neural
network



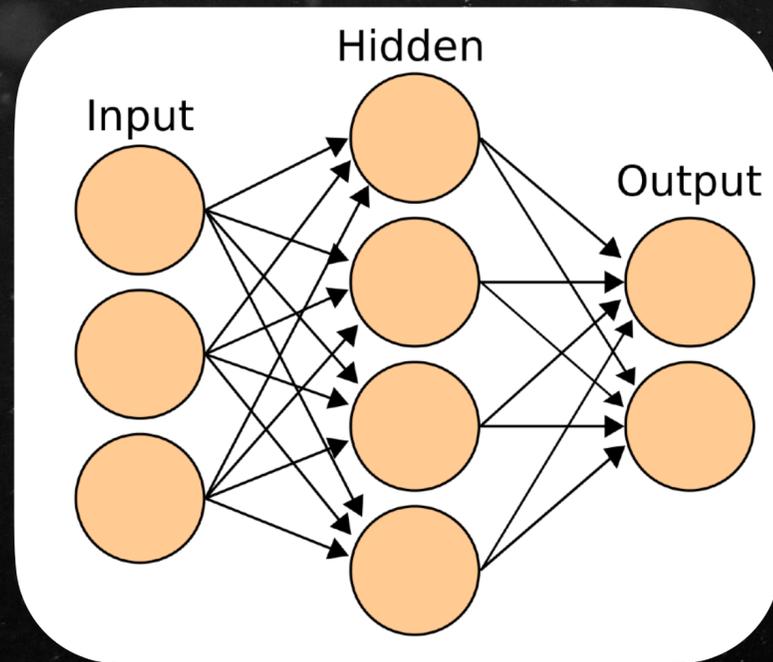
The Learning Boundary



Next Gen: Interpretable AI

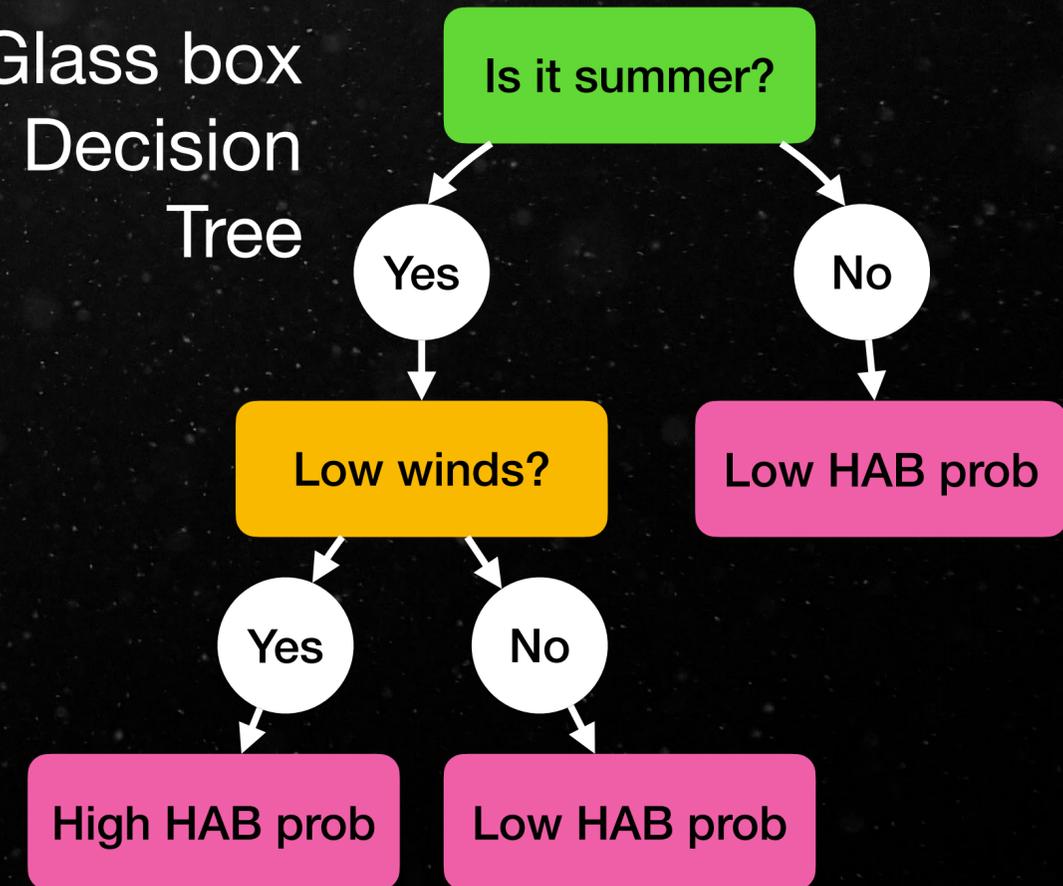
Machine learning (e.g. neural nets) offer accurate predictions, but you don't know what's going on under the hood. **Interpretable AI** is a class of machine learning that people can understand

Black box
neural
network



The Learning Boundary

Glass box
Decision
Tree



Next Gen: Transfer Learning

Water quality modeling suffers from the “lots of small data” problem (i.e. its not a big data problem).



Next Gen: Transfer Learning

Water quality modeling suffers from the “lots of small data” problem (i.e. its not a big data problem).

Real-time
Data



Lots of
Historical
Data



Next Gen: Transfer Learning

Water quality modeling suffers from the “lots of small data” problem (i.e. its not a big data problem).

Real-time
Data



Lots of
Historical
Data



Real-time
Data

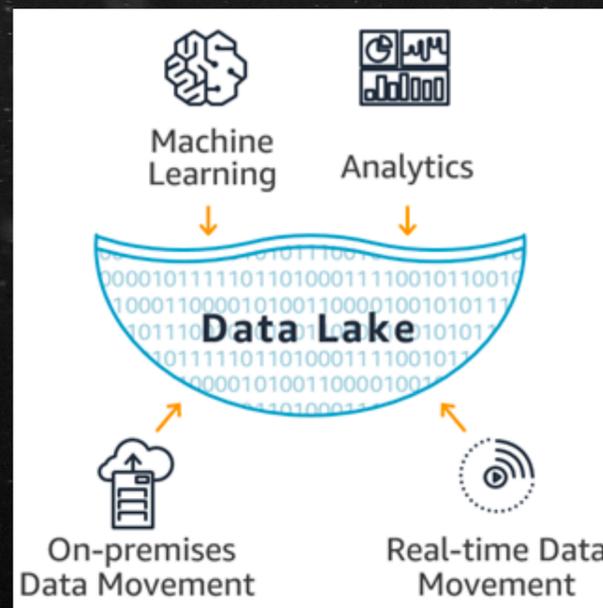


Little
Historical
Data



Cloud Data Infrastructure

It's not just about modeling though. We're learning how important it is to have a unified, standardized and accessible (via API) database: AWS Data Lake + database...



Unified, standardized Data Lake and database



Collaborators in a 2021 Water Research Foundation project to develop a national water quality database for application to transfer learning



Conclusions: Measurable Usefulness

In sum, collectively we are doing really well monitoring and modeling our environments. I think the next steps are:

- 1) **synthesize all available data in a standardized database**
(Cyan, water samples, qPCR, cubesat imagery, USGS... etc etc)
- 2) **developing metrics** by which we can measure the impact of our monitoring efforts and modeling.

In Summary:

- Multimodal data synthesis and open access
- Models: interpretable AI and transfer learning
- Predictions: at weekly, monthly and decadal timescales
- Measurable impacts



LAke multi-scaled GeOSpatial and temporal database

Publicly accessible lake water quality and ecological context data for the US

University of Michigan





The Prediction Lab

james@thepredictionlab.com



T
P
L